5-1-2020

A Clustering Algorithm for Early Prediction of Controversial Reddit Posts

Abenezer Daniel Dara
Dartmouth College

Dartmouth Computer Science Technical Report TR2020-891

A CLUSTERING ALGORITHM FOR EARLY PREDICTION OF CONTROVERSIAL REDDIT POSTS

A Thesis

Submitted to the Faculty

in partial fulfillment of the requirement for the

degree of

Bachelor of Arts

in

Computer Science

By

Abenezer Daniel Dara

Advisor: Professor Soroush Vosoughi

DARTMOUTH COLLEGE

Hanover, New Hampshire

May 2020

Abstract

Social curation platforms like Reddit are rich with user interactions such as comments, upvotes, and downvotes. Predicting these interactions before they happen is an interesting computational challenge and can be used for a variety of tasks, ranging from content moderation to personality prediction. Given the vast amount of information posted on these sites, it's important to develop models that can simplify this prediction task. In this paper, we present a simple clustering algorithm that helps predict the controversiality of a Reddit post using the user's profile information, their past contributions on Reddit, and the sentiment expressed in their post. On average, introducing the cluster to the prediction task improved the accuracy of the prediction by over 20 percent, with F1 scores of 0.95 (micro) and 0.7 (macro). The classifier performs better than a majority predictor. The results also show that the overwhelming majority of users are inactive and when they do post, they post non-controversial content.

Contents

Chapter 1

Introduction

Section 1.1

Problem Statement

Social media sites have become common opinion sharing platforms (Addawood and Bashir 2016). As a result, academics have long been interested in understanding the interactions between users on social media sites. From personality prediction to understanding how misinformation spreads online, understanding the interaction between social media users has been informative to researchers (Hessel et al. 2019; Addawood et al. 2016; Angeletou et al. 2011). In this paper, we present a clustering algorithm that helps detect the controversiality of posts on Reddit, a social news aggregation discussion web site.

Detecting these controversial posts has several advantages. Prior research has shown that the controversiality of posts can be indicative of anti-social behavior by the individual posting them (Smith et al. 2013). Detecting these posts early before they receive any comments can help identify content that needs to be moderated (Morrison and Hayes 2013). It can also be used to create a system where the social media site can warn the user when they are about to post a controversial content,

giving them a chance to modify their post if that was not their intent. Furthermore, detecting controversiality can also be used to identify discussion threads where debate is happening. This can be used to sort "hot" discussion threads and engage the online community in a discussion. In addition, detecting controversial posts can also help auto-detect bullying behavior (Medvedev et al. 2017).

However, predicting the responses a post receives is an expensive and complicated computational task (Burlutskiy et al. 2016). Given the large volume of user activity on platforms such as Reddit, training models to recognize a pattern of activity is an expensive computational task (Burlutskiy et al. 2016; Lim et al. 2017). On top of that, we also need to account for context-specific variables. For instance, whether a particular post will be controversial or not may depend on the subreddit the post is made instead of the content of the post. For instance, posts on the topic of breakups are controversial on the relationships subreddit, while they do not generate controversy in the AskWomen subreddit (Hessel and Lee 2019). Moreover, other researchers have shown that the controversiality of a post is determined by the first few comments it receives instead of the original content of the post (Chang and Mizil 2019; Hessel and Lee 2019).

This paper presents a simple clustering algorithm for the detection of controversial posts well before they receive any responses from the subreddit community. The algorithm is trained on data from over 700,000 Reddit users in 50 subreddits for the period spanning 2004-2018. We selected Reddit because it is one of the most commonly used social media networks and ranks in the top 10 most visited websites in the US [1]. And since signing up for Reddit does not require an email address, it attracts a large number of users, giving us more content to work with. Because all posts on Reddit are public, we can download all submissions made within the

[1] alexa.com/siteinfo/reddit.com

timeframe of our interest as long as they are not deleted. Reddit is divided into "communities" known as subreddits, where users talk about a specific topic related to the subreddit. For this paper, we will focus on 50 subreddits where Redditors discuss economics and finance [2].

Section 1.2

Controversial Posts

Prior researchers have defined controversial posts as those that cause polarization, receiving both significant positive and negative comments (Hessel and Lee 2019). These posts are more likely to invited heated debate, attracting responses that span a wide variety of emotions. The controversiality of a post could be measured in several ways.

One common method is to examine the sentiment expressed in the post and the comments it attracts. In several studies, researchers have used sentiment analysis to predict if comments for a post are supportive, neutral, or against the post (Smith et al 2013). The researchers used these categorizations of comments to calculate the controversiality of the post. However, this method fails to capture controversiality when it is not expressed in words. Upvotes and downvotes a post receives are also strong indicators of "community" opinion, but this method fails to capture them. On top of that, many replies to a post can be links to outside sites or memes that do not lend themselves easily for sentiment analysis programs.

As a result, we need to take into account upvotes and downvotes to measure

[2]Subreddits examined: r/finance,r/economy, r/AskEconomics, r/jobs, r/workonline,r/forhire, r/PersonalFinance, r/Entrepreneur, r/startups, r/financialindependence, r/realestate, r/flipping, r/antimlm, r/ripple, r/Iota, r/stellar, r/investing, r/wallstreetbets, r/millionairemakers, r/weedstocks, r/frugal, r/EatCheapAndHealthy,r/frugalmalefashion, r/budgetfood, r/cheap_meals,r/Frugal_Jerk, r/povertyfinance, r/shutupandtakemymoney, r/BuyItForLife, r/crappyoffbrands, r/shouldibuythisgame, r/Anticonsumption, r/sbubby, r/Wellworn, r/ineeeedit, r/didntknowiwantedthat, r/Bitcoin, r/dogecoin, r/CryptoCurrency, r/ethereum, r/ethtrade, r/litecoin, r/btc, r/garlicoin, r/cardano, r/Vechain

the controversiality of posts. However, Reddit no longer provides upvote/downvote counts for posts. Instead, it provides the upvote and downvote count for comments made under posts. We will use these to calculate the controversiality score for each comment and use it as a proxy to measure the controversiality of a post. The formula for controversiality is given by $(u+d)^{min(\frac{u}{d},\frac{d}{u})}$, where u and d correspond to the number of upvotes and downvotes respectively.

The data we obtained shows that the vast majority of posts are not controversial. There are two main reasons for this. First, the vast majority of posts on Reddit do not get any replies. And second, the Reddit algorithm shows the most popular content to users on the first pages of the subreddits. Because controversial posts are not categorized in the popular category, they are made less visible to users (Morrison and Hayes 2013).

Chapter 2

Data and Methodology

Section 2.1

Data

We downloaded posts and comments from 50 subreddits spanning the period between 2004 and 2018. The data was obtained from Google's BigQuery database [1]. User account data was obtained using Reddit's PRAW API. Our dataset contains 718,732 users.

2.1.1. Organizing the Data

To organize the data, we represented each user as a 58-dimensional vector. The first eight dimensions contained the information we collected about each user using Reddit's PRAW API. The fields are given below [2]:

[1] https //console.cloud.google.com/bigqueryproject=fhbigquery&p=fh-bigquerydreddit&pagedataset

[2] https://praw.readthedocs.io/en/latest/code_overview/models/redditor.html

Field	Description
created_utc	Unix timestamp of when the account was created
verified_email	Whether or not the Redditor has verified their email.
is_reddit_employee	Whether or not the Redditor is a Reddit employee.
is_mod	Whether or not the Redditor is the moderator of any subreddit.
is_gold_user	Whether or not the Redditor has active Reddit Premium Status
link_karma	The score of how popular the Redditor's posts are. This is akin to a reputation score of the user.
comment_karma	The score of how popular the Redditor's comments are.
hasPrawData	Indicates whether or not the account has been deleted by Reddit.

Table 2.1: Data collected for each user

Some accounts from the initial dataset were deleted. As a result, we could not obtain any of the above data for them. To account for this, we added another feature (hasPrawData) that indicates whether or not we were able to download the user's information. The next 50 dimensions captured the user's contribution to the subreddits we are working with. Here, each field represented the total number of times the user commented or posted in each of our subreddits. Prior research has shown that user contribution captures important qualities of Redditors such as their level of expertise on the topics discussed in their subreddit communities (Lim et al. 2017).

To reduce the dimension of the data, we employ Principal Component Analysis (PCA), an unsupervised feature extraction method that allows us to reduce the dimension of a given data by identifying a smaller number of variables that summarize our large dataset [3]. By initializing a PCA with 58 components, we observe that the first two components explain more than 99.99% of the covariance in our data. Consequently, using a two-component PCA, we reduce our data to be two-dimensional.

[3]https://www.sciencedirect.com/topics/medicine-and-dentistry/principal-component-analysis

Section 2.2

Clustering

Once we obtained our two-dimensional data, we used a K-means classifier to cluster the users into separate buckets. K-means is an unsupervised classification algorithm that partitions a given data into k distinct, non-overlapping clusters. These clusters minimize the sum of squared distances between data points and the mean of the data points in the cluster while maximizing the distance between the mean of each cluster [4]. We then performed the elbow method to determine the number of clusters for a K-means clustering (k=3) on the two-dimensional data we created. The clustering partitioned the users into three buckets that represented 37%, 10%, and 53% of the users in the initial dataset.

The clustering partitioned the users into the following three buckets:

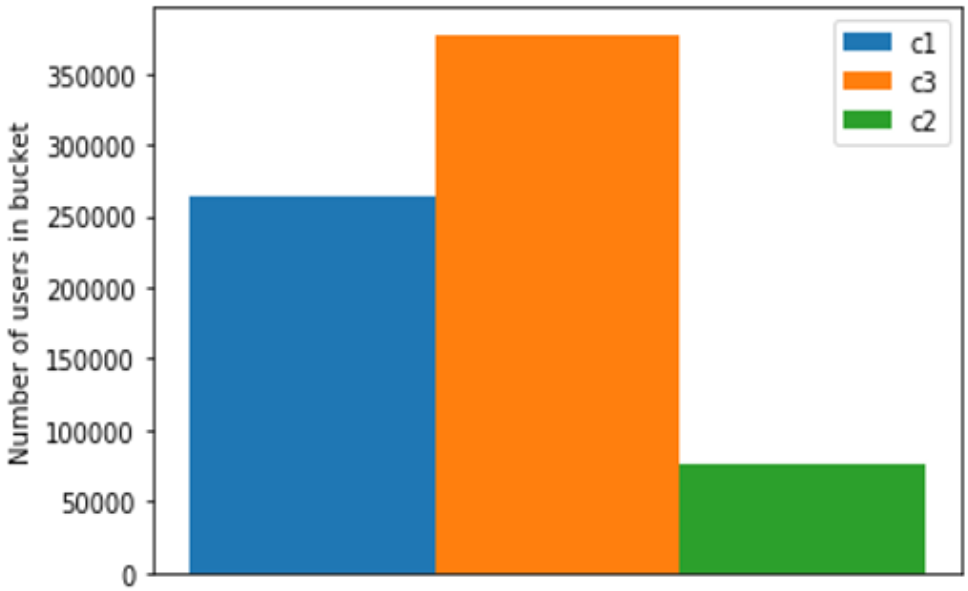

Figure 2.1: Number of users per cluster

[4] https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

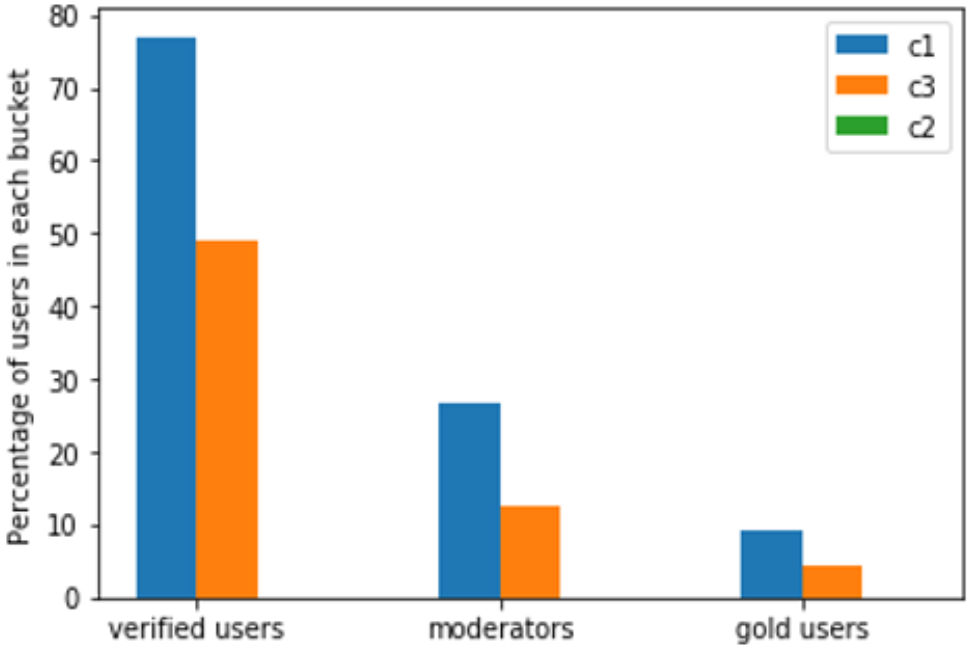

Figure 2.2: Percentage of users in each bucket by user feature

The Clusters. The descriptive statistics show that Redditors in the first cluster are the vast majority of the users who have verified their email, have a high reputation (karma) score, and are moderators or gold users. On the other hand, the second cluster contains Redditors whose accounts have been deleted or suspended by Reddit. Even though their accounts were deleted or suspended, their contributions represented about 10% of our dataset, therefore we decided to keep them in our analysis. Finally, the third cluster contains the rest of the users, and this is where the majority of the users fall.

Using the same method as above, we further partitioned each bucket into three clusters. We reduced the dimensionality using top Principal Component Analysis that accounted for 99.9% of the variance in the data. We then run K-means (k=3) to obtain nine clusters for the entire user data.

Section 2.3

Prediction Task

After obtaining the controversiality score for each post, we turn to the prediction task. To detect the sentiment expressed in each post, we used Vader sentiment analysis that is specifically attuned to sentiments expressed in social media posts and comments. We obtain the scores for how negative, positive, or neutral a certain text is. For each post in our dataset, we used Vader to calculate a single compound score between -1 (very negative) and 1 (very positive) indicating the intensity and direction of the sentiment expressed in the post. Posts that did not contain any words were not included in this step.

Reddit no longer provides the number of downvotes on posts. Therefore, to detect the controversiality of posts for our training data, we measured the controversy generated in the comments section as a proxy. We were able to obtain the upvotes and downvotes for each comment, which allowed us to calculate the controversiality score using the following formula: $(u + d)^{min(\frac{u}{d}, \frac{d}{u})}$, where u and d correspond to the number of upvotes and downvotes respectively.

Chapter 3

Results

For our baseline, we train a Gaussian Naïve Bayes (GNB) classifier to predict the controversiality score of the Reddit posts using the sentiment scores from Vader. GNB is a variation of the Naïve Bayes classification algorithm that can be applied to continuous and normally distributed data. The algorithm computes conditional class probabilities and predicts the most likely value of the target feature, in our case the controversiality score [1]. For our second prediction task, we add the nine user clusters obtained from our previous computation and retrained the Gaussian model. We then measured if adding the buckets improves the quality of our prediction.

Our results show an improvement in F1 scores with the inclusion of the nine user clusters. On average, we recorded an improvement in the Macro F1 score from 0.5 to 0.7, and the Micro F1 improved from 0.86 to 0.95. The Micro F1 score starts from a higher value because of the imbalance in the data; The overwhelming majority of posts in our dataset ($> 90\%$) have very low controversiality scores and received fewer than 10 comments.

[1] https //towardsdatascience.com/naive-bayesclassifier-explained-50f9723571ed

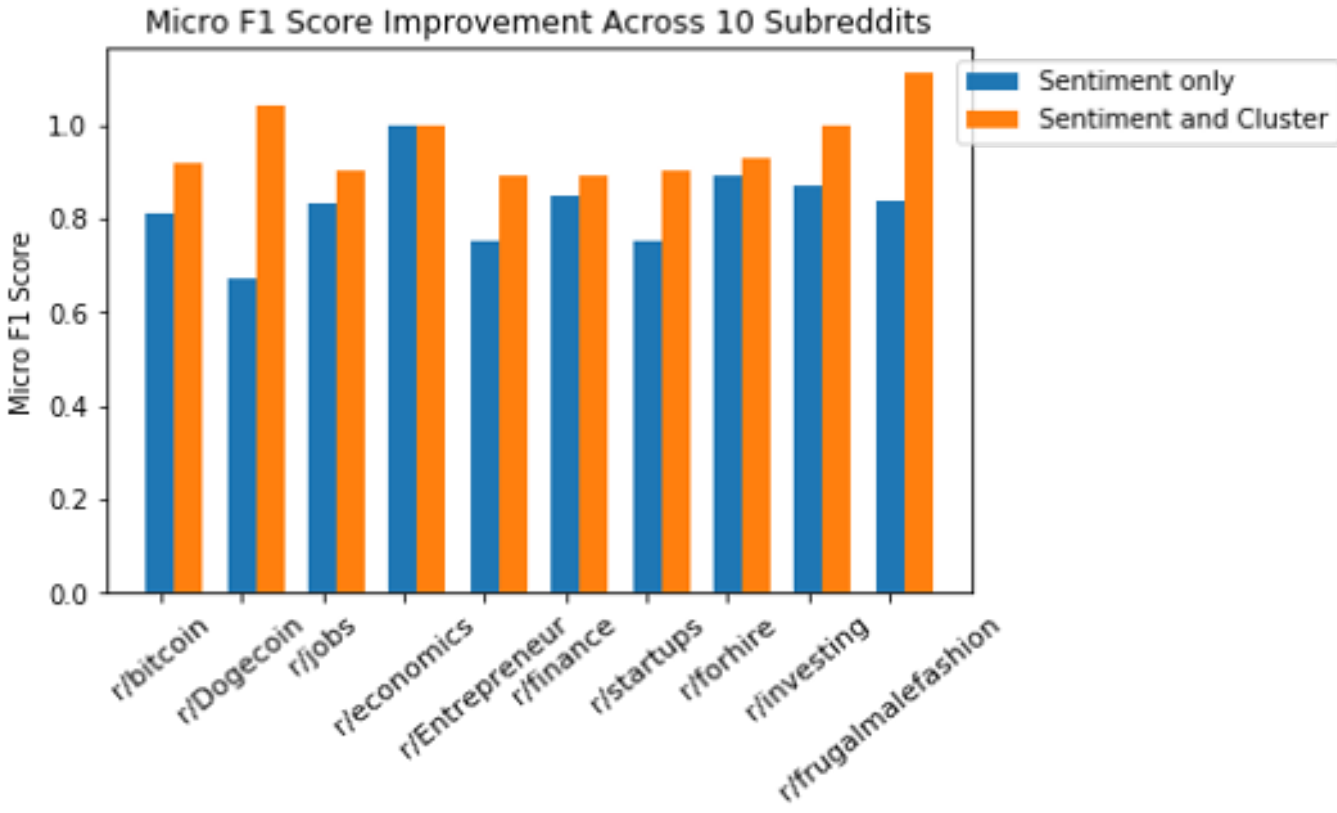

Figure 3.1: Micro F1 Score Improvements before and after including the user clusters

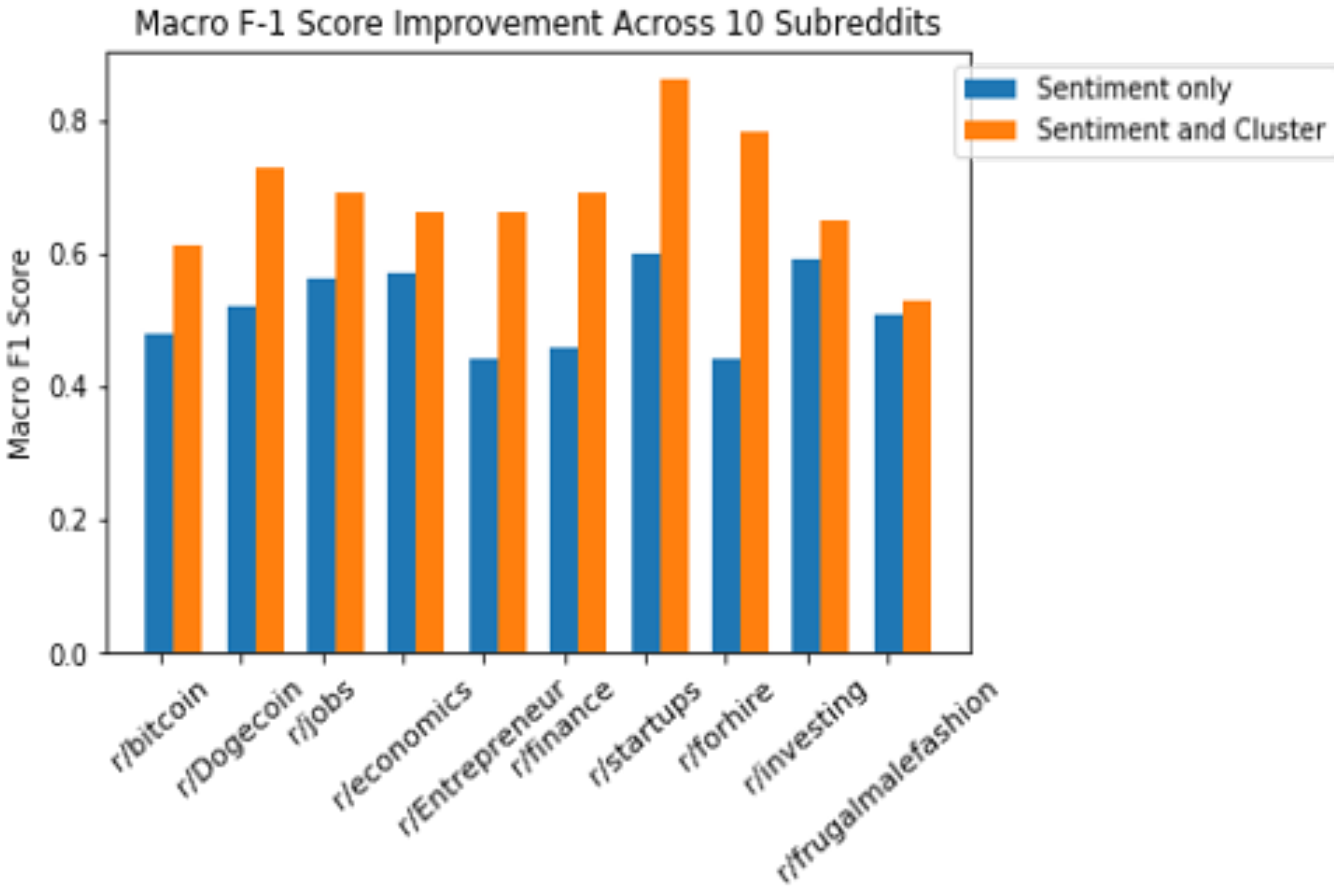

Figure 3.2: Macro F1 Score Improvements before and after including the user clusters

Chapter 4

Discussion

Section 4.1

Limitation and Future Work

We show how to more accurately predict the controversiality of Reddit posts by clustering users into different buckets based on their prior contribution to Reddit and basic information from their user accounts. Including these user clusters in our prediction task improved the F1 score of our controversiality prediction, indicating that the clusters capture important qualities about the users. Further work can explore other characteristics that these clusters capture. For instance, predicting the popularity of users or predicting posts that are likely to be ignored.

However, the work has several limitations that future work needs to address. Many posts are links to other websites or just images which do not lend themselves easily to sentiment analysis. Those posts were not used in our model due to this limitation. In addition, further work needs to account for the technical challenges of mapping comment replies to the original posts when the parent comments are deleted. We had to exclude these replies from our dataset. Additionally, increasing the number of buckets could improve the prediction quality and needs to be explored.

Bibliography

Addawood, Aseel, and Masooda Bashir. “What is Your Evidence?” A Study of Controversial Topics on Social Media." Proceedings of the Third Workshop on Argument Mining (ArgMining2016). 2016.

Alexa. https://www.alexa.com/siteinfo/reddit.com#section_traffic

Angeletou, Sofia, Matthew Rowe, and Harith Alani. "Modelling and analysis of user behaviour in online communities." International Semantic Web Conference. Springer, Berlin, Heidelberg, 2011.

Burlutskiy, Nikolay, et al. "An investigation on online versus batch learning in predicting user behaviour." International Conference on Innovative Techniques and Applications of Artificial Intelligence. Springer, Cham, 2016.

Chang, Jonathan P., and Cristian Danescu-Niculescu-Mizil. "Trouble on the Horizon: Forecasting the Derailment of Online Conversations as they Develop." arXiv preprint arXiv:1909.01362 (2019).

Hessel, Jack, and Lillian Lee. "Something's Brewing! Early Prediction of Controversy-causing Posts from Discussion Features." arXiv preprint arXiv:1904.07372 (2019).

Lim, Wern Han, Mark James Carman, and Sze-Meng Jojo Wong. "Estimating relative user expertise for content quality prediction on Reddit." Proceedings of the 28th ACM Conference on Hypertext and Social Media. 2017.

Medvedev, Alexey N., Renaud Lambiotte, and Jean-Charles Delvenne. "The anatomy of Reddit: An overview of academic research." Dynamics on and of Complex Networks. Springer, Cham, 2017.

Morrison, Donn, and Conor Hayes. "Here, have an upvote: Communication behaviour and karma on Reddit." INFORMATIK 2013–Informatik angepasst an Mensch, Organisation und Umwelt (2013).

Smith, Laura M., et al. "The role of social media in the discussion of controversial topics." 2013 International Conference on Social Computing. IEEE, 2013.